

Copyright  
by  
Max Edward Robinson  
2013

**The Thesis Committee for Max Edward Robinson**  
**Certifies that this is the approved version of the following thesis:**

**Structural investigations of the group II intron-encoded protein GsI-IIC**

**APPROVED BY**  
**SUPERVISING COMMITTEE:**

**Supervisor:**

---

Alan Lambowitz

---

Rick Russell

**Structural investigations of the group II intron-encoded protein GsI-IIC**

**by**

**Max Edward Robinson, BS**

**Thesis**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Master of Arts**

**The University of Texas at Austin**

**August 2013**

## **Abstract**

### **Structural investigations of the group II-intron encoded protein GsI-IIC**

Max Edward Robinson, MA

The University of Texas at Austin, 2013

Supervisor: Alan Lambowitz

Group II introns are a class of mobile ribozymes found in bacteria and eukaryotic organelles that self-splice from precursor RNAs. The resulting lariat intron RNA can then insert into new genomic DNA sites through a reverse splicing reaction. Collectively, this process of intron mobility is termed “retrohoming.” Mobile group II introns encode a reverse transcriptase (RT) that stabilizes the catalytically active form of the intron RNA for both the forward and reverse splicing reactions and also converts the integrated intron RNA into DNA. This work aims to elucidate the structure of bacterial group II intron-encoded RTs and ultimately determine how they function in intron mobility. Although efforts to crystallize group II introns RTs have been unsuccessful, small angle X-ray scattering studies in conjunction with homology modeling have provided new insights into the structure and function of these enzymes.

## Table of Contents

List of Tables .....	vii
List of Figures .....	viii
Chapter 1: Introduction .....	1
Mobile group II introns .....	1
Mechanism of group II intron mobility .....	1
Group II intron splicing and RNA structure .....	3
Group II intron-encoded proteins .....	5
Group II introns from <i>Geobacillus stearothermophilus</i> .....	7
Overview of thesis research .....	9
Chapter 2: Experimental Design and Methods .....	11
Sample preparation .....	11
Expression of group II intron-encoded proteins .....	11
Cloning of expression plasmids .....	12
Protein expression and purification .....	13
Group II intron ribonucleoprotein preparation .....	15
<i>In vitro</i> synthesis of group II intron precursor RNA .....	15
Cloning of RNA template plasmids .....	17
Synthetic substrate mimics .....	18
X-ray crystallography .....	19
Surface entropy reduction .....	20
Small angle X-ray scattering .....	21
Molecular modeling .....	25
Chapter 3: Results and Discussion .....	28
Sample preparation .....	28
X-ray crystallography .....	29
Small angle X-ray scattering .....	31
Molecular modeling .....	33

References .....	36
------------------	----

## **List of Tables**

Table 1: Mutagenic primers used for generating pSER-GsI-IIC constructs .....	21
Table 2: GsI-IIC SER constructs .....	30
Table 3: MRF-GsI-IIC structural parameters calculated from SAXS data.....	32
Table 4: Hydrodynamic and gyration radii .....	32

## List of Figures

Figure 1:	Group II intron retrohoming pathway in bacteria .....	3
Figure 2:	Group II intron splicing mechanism.....	4
Figure 3:	Group II intron RNA secondary structure .....	5
Figure 4:	Bacterial group II intron-encoded proteins .....	7
Figure 5:	Phylogeny of group II intron ORFs and correspondence with RNA structural classes .....	8
Figure 6:	Schematic of the group II intron encoded-protein GsI-IIC from <i>Geobacillus stearothermophilus</i> .....	9
Figure 7:	Schematic of MBP-GsI-IIC fusion proteins .....	12
Figure 8:	Model of the GsI-IIC intron and schematic of the precursor RNA construct used for RNP complex formation.....	18
Figure 9:	Synthetic substrates mimics .....	19
Figure 10:	SAXS experimental set up and data collection .....	23
Figure 11:	The distance distribution function.....	24
Figure 12:	Schematic representation of the I-TASSER protocol.....	27
Figure 13:	Gel filtration chromatogram .....	29
Figure 14:	MRF-GsI-IIC SAXS scattering profile and shape reconstruction .....	32
Figure 15:	Homology model of GsI-IIC .....	33
Figure 16:	Predicted orientation of MBP and GsI-IIC in MRF-GsI-IIC .....	34
Figure 17:	<i>T. castaneum</i> telomerase catalytic subunit .....	35



## Chapter 1: Introduction

### MOBILE GROUP II INTRONS

Group II introns are self-splicing RNAs (“ribozymes”) found in bacteria and eukaryotic organelles that act as mobile retroelements [Cech 1986 and Cavalier-Smith 1991]. Characteristics of group II introns, including their structure and splicing mechanism, suggest that they are evolutionary ancestors of eukaryotic spliceosomal introns and the spliceosome itself [Lambowitz and Zimmerly 2004 and Lambowitz and Zimmerly 2010]. Group II introns fold into a highly conserved three-dimensional structure containing six distinct helical domains [Michel *et al.* 2009]. The majority of bacterial group II introns contain an open reading frame (ORF) that encodes a multifunctional protein with an N-terminal reverse transcriptase (RT) domain that shares homology with retroviral RT sequences [Blocker *et al.* 2005]. Additionally, the intron-encoded protein (IEP) contains a putative RNA-binding domain that corresponds to the RT thumb and is implicated in RNA folding and splicing (“maturase”) activities [Wank *et al.* 1999 and Mohr *et al.* 1993]. Certain classes of IEPs contain a C-terminal DNA-binding domain, and others have an additional DNA endonuclease domain [San Filippo and Lambowitz 2002]. After splicing, the intron RNA and the IEP function together in a ribonucleoprotein (RNP) complex to promote intron mobility.

### MECHANISM OF GROUP II INTRON MOBILITY

Most of what is known about group II intron retrohoming has come from studies of the *Lactococcus lactis* Ll.LtrB model system [Yao *et al.* 2013]. Figure 1 diagrams the key steps in the retrohoming pathway of a generalized bacterial group II intron. The transcribed group II intron precursor RNA consists of the intron sequence flanked by 5'- and 3'-exons (E1 and E2, respectively). The intron-encoded protein (denoted IEP) is

translated from within the intron and is a multifunctional protein with RT, RNA splicing, DNA binding, and DNA endonuclease activities [Matsuura *et al.* 1997 and Saldanha *et al.* 1999]. The IEP promotes splicing by binding to the intron in the unspliced precursor RNA and stabilizing the catalytically active RNA structure [Matsuura *et al.* 2001]. Splicing occurs via two sequential transesterification reactions that are catalyzed by the intron RNA, yielding ligated exons and an excised intron lariat RNA. The IEP remains tightly bound to the intron lariat in an RNP that can initiate retrohoming by recognizing an appropriate DNA target sequence (ligated E1-E2 DNA sequence) via the IEP and base pairing of the intron RNA [Guo *et al.* 2000, Mohr *et al.* 2000, and Perutka *et al.* 2004]. The intron RNA then inserts between the two DNA exons by reverse splicing into the top strand of the DNA, and the IEP cleaves the bottom strand downstream of the intron-insertion site. After cleavage, the IEP uses the 3' end of the DNA strand for target DNA-primed reverse transcription (TPRT) of the newly inserted RNA [Matsuura *et al.* 1997 and Saldanha *et al.* 1999]. Integration of the resulting intron cDNA is carried out by host factors in late steps that include RNA degradation, top-strand DNA synthesis, resection of DNA overhangs, and DNA ligation [Yao *et al.* 2013].

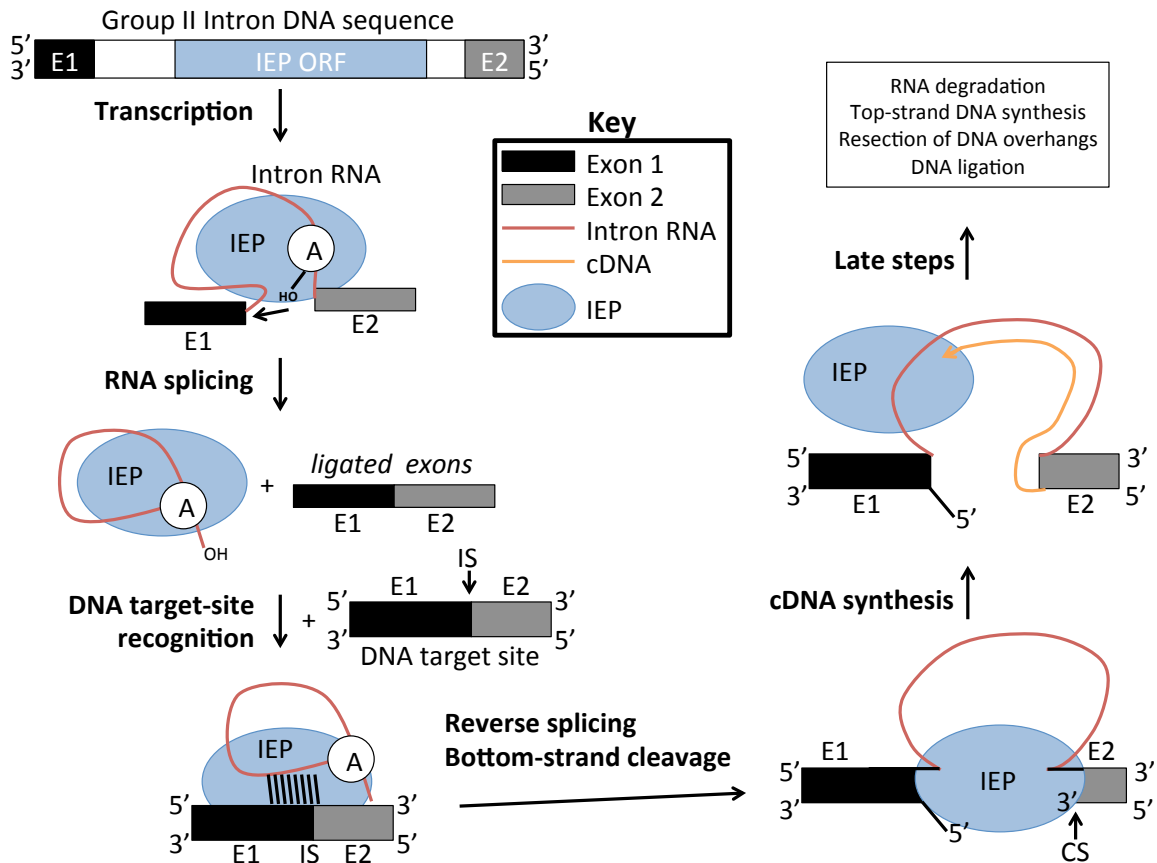


Figure 1: Group II intron retrohoming pathway in bacteria. The group II intron-encoded protein (IEP) and the spliced intron RNA function together as a ribonucleoprotein (RNP) that recognizes and binds a DNA target-site. The intron RNA reverse splices into the top strand of the DNA insertion site (IS), and the IEP cleaves the bottom strand at the cleavage site (CS) to allow for DNA-primed reverse transcription of the integrated RNA. Host factors are responsible for late steps that include RNA degradation and top-strand DNA synthesis. Adapted from Yao *et al.* 2013.

## GROUP II INTRON SPLICING AND RNA STRUCTURE

Like eukaryotic spliceosomal introns, group II introns splice via two sequential transesterification reactions that yield ligated exons and an excised intron lariat with a 2'-5' phosphodiester bond (Figure 2) [Lambowitz and Zimmerly 2004]. In contrast with spliceosomal introns, the group II intron splicing reaction is catalyzed by the intron RNA itself [Lambowitz and Zimmerly 2004]. To allow for splicing, the RNA folds into

conserved secondary and tertiary structures that form an active site containing  $Mg^{2+}$  ions essential for catalysis [Lambowitz and Zimmerly 2004]. The active site binds the splice sites and the branch-point nucleotide A and uses  $Mg^{2+}$  ions to activate the appropriate bonds for catalysis [Lambowitz and Zimmerly 2010].

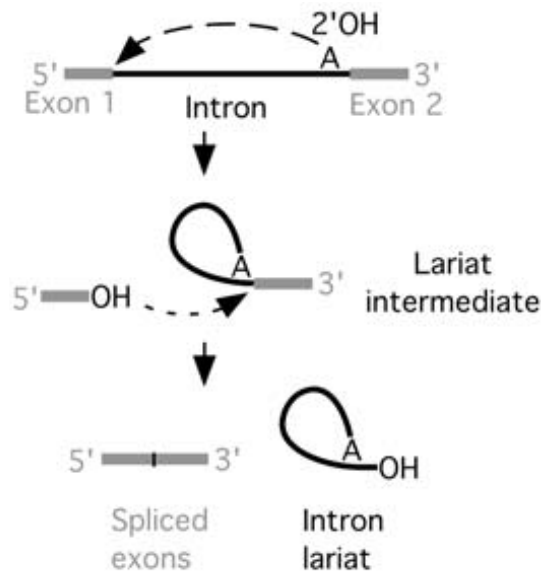


Figure 2: Group II intron splicing mechanism. Group II intron splicing is catalyzed by the RNA itself and occurs via two sequential transesterification reactions. In the first reaction, there is a nucleophilic attack of the 5'-splice site by the 2' OH of a bulged A residue in DVI, resulting in cleavage of the 5'-splice site and the lariat intermediate. In the second reaction, there is a nucleophilic attack of the 3'-splice site by the 3' OH of the cleaved 5' exon, yielding ligated exons and the intron lariat with a 2'-5' phosphodiester bond. Adapted from Lambowitz and Zimmerly 2004.

Group II intron RNA secondary structure is organized into six helical domains DI through DVI radiating outward from a central wheel (Figure 3) [Keating *et al.* 2010]. These domains fold into a conserved tertiary structure that brings together distant sequences in DI and DV to form the catalytic core [Lambowitz and Zimmerly 2010]. DV contains the catalytic triad, DVI contains the branch-point nucleotide (a bulged A residue), and DII and DIII contribute to RNA folding and catalysis [Keating 2010 and

Lambowitz and Zimmerly 2010]. DIV encodes the intron ORF not necessary for catalysis and also contains a high-affinity binding site for the IEP near its 5' end [Lambowitz and Zimmerly 2010]. Critical for folding are the short exon-binding sequences EBS1/2 in DI that form base pairs with the intron-binding sequences IBS1/2 in the 5'-exon and the  $\delta$  sequence in DI that forms base pairs with the  $\delta'$  sequence in the 3'-exon [Lambowitz and Zimmerly 2004].

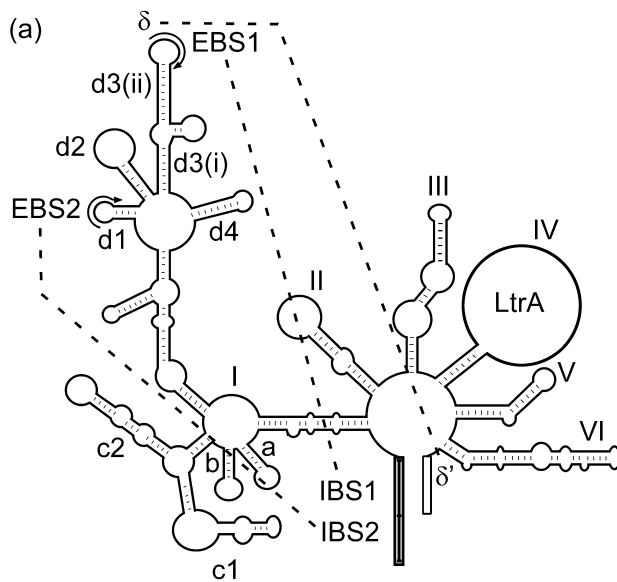


Figure 3: Group II intron RNA secondary structure: model of the *L. lactis* LI.LtrB intron. The predicted secondary structure consists of six double-helical domains I-VI radiating out from a central wheel. Dashed lines represent long-range interactions between the intron and exons in the unspliced precursor RNA (EBS1-IBS1, EBS2-IBS2, and  $\delta$ - $\delta'$ ). The ORF in DIV encodes the LtrA protein. Adapted from Perutka *et al.* 2004.

## GROUP II INTRON-ENCODED PROTEINS

Although group II intron RNA self-splicing can be observed *in vitro*, this reaction requires non-physiological conditions [Lambowitz and Zimmerly 2004]. *In vivo*, the group II intron-encoded protein is required to help the intron fold into a catalytically active structure (“maturase” activity) [Lambowitz and Belfort 1993]. To stabilize the

active structure, the group II intron IEP binds specifically to the intron RNA, with this binding interaction likely mediated by residues in the RT and X (thumb) domains of the IEP (Figure 4) [Lambowitz and Zimmerly 2004].

The group II intron ORF encoded within DIV of the intron RNA consists of four distinct domains denoted RT, X, DNA binding (D), and DNA endonuclease (En) (Figure 4) [Lambowitz and Zimmerly 2004]. The N-terminal RT domain contains conserved sequence blocks RT-1 through RT-7, which correspond to the fingers and palm of retroviral RTs, and an upstream RT-0 sequence characteristic of non-long terminal repeat (LTR) retrotransposons RTs (Figure 4) [Blocker *et al.* 2005]. Domain X is located just downstream of the RT domain in a position that corresponds to the thumb domain of retroviral RTs and has been implicated in maturase activity [Cui *et al.* 2004]. The RT and X domains function together to bind the intron RNA as a substrate for splicing and as a template for reverse transcription [Cui *et al.* 2004]. Based on evidence from molecular modeling studies of LtrA, it has been suggested that the evolutionarily conserved “insertion” sequences (relative to retroviral RTs) in the RT domain are important for mediating interactions with intron RNA [Blocker *et al.* 2005]. The C-terminal D and En domains are not necessary for RNA splicing. However, these domains do function in intron mobility [Lambowitz and Zimmerly 2004]. Domain D is not highly conserved in sequence, but contains a cluster of basic amino acid residues and a predicted  $\alpha$ -helix, both of which are functionally important features [San Filippo and Lambowitz 2002]. The En domain is responsible for second-strand DNA cleavage during retrohoming and consists of motifs characteristics of the H-N-H DNA endonuclease family [San Filippo and Lambowitz 2002].

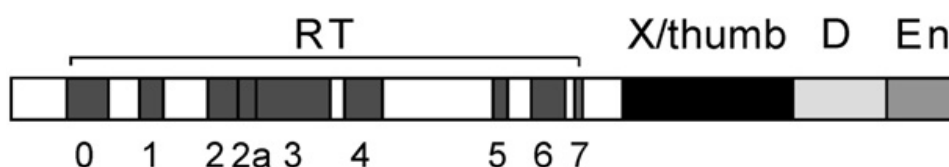


Figure 4: Bacterial group II intron-encoded proteins: schematic of the group II intron-encoded protein LtrA from *L. lactis*. LtrA is a multifunctional protein with four distinct domains denoted reverse transcriptase (RT), X/thumb, DNA binding (D), and DNA endonuclease (En). The N-terminal RT domain consists of conserved sequence blocks 0-7. Adapted from Dai *et al.* 2008.

#### GROUP II INTRONS FROM *GEOBACILLUS STEAROTHERMOPHILUS*

The thermophilic bacterium *Geobacillus stearothermophilus* was recently found to contain 17 copies of a group II intron present at different locations along its genome [Moretz and Lampson 2010]. These introns belong to the bacterial group IIC subclass (Figure 5) and encode a C-type IEP that lacks an N-terminal En domain [Moretz and Lampson 2010 and Lambowitz and Zimmerly 2010]. The GsI-IIC ORF (Figure 6) corresponding to one of these genomic sequences encodes a 420-amino-acid, heat-stable protein with reverse transcriptase activity referred to as GsI-IIC [Vellore *et al.* 2004]. Additionally, a GsI-IIC fusion protein has been shown to support retrohoming *in vivo* in an *E. coli* plasmid-based assay developed by our lab [Mohr *et al.* 2013]. Furthermore, GsI-IIC has been predicted to have a relatively high crystallization probability compared to other group II IEPs based on its inherent biochemical and biophysical characteristics (as determined by the XtralPred-RF server) [Slabinski *et al.* 2007]. For these reasons, GsI-IIC was selected as a candidate for an in depth structural investigation of group II IEPs.

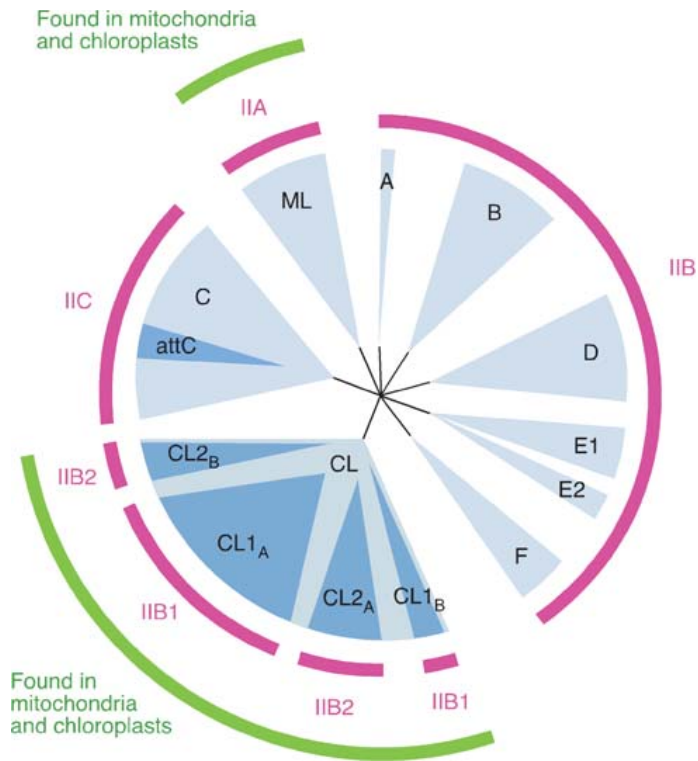


Figure 5: Phylogeny of group II intron ORFs and correspondence with RNA structural classes. The major lineages of group II intron IEPs (blue sectors) are denoted chloroplast-like (CL), mitochondrial like (ML), and bacterial classes (A-F). The RNA structural subclasses that correspond to IEP lineages are shown in magenta. All group II lineages can be found in bacteria, and those also found in eukaryotic organelles are shown in green. The group II introns found in *G. stearothermophilus* are of the RNA structural subclass IIC and they encode a C-type IEP. Adapted from Lambowitz and Zimmerly 2010.



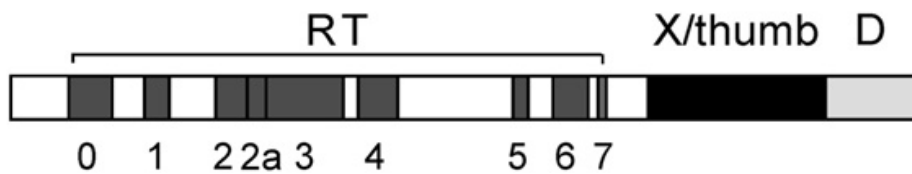


Figure 6: Schematic of the group II intron encoded-protein GsI-IIC from *Geobacillus stearothermophilus*. Like LtrA, GsI-IIC is a multifunctional protein with an N-terminal reverse transcriptase (RT) domain that consists of conserved sequence blocks 0-7. GsI-IIC also contains X/thumb and DNA-binding (D) domains, but lacks a C-terminal endonuclease (En) domain. Adapted from Dai *et al.* 2008.

## OVERVIEW OF THESIS RESEARCH

This work aims to develop a structural framework for understanding group II intron mobility. Specifically, I seek to provide a detailed description of how group II intron-encoded proteins promote intron self-splicing and mobilization to new genomic DNA sites. It has been hypothesized that evolutionarily conserved insertion sequences in the RT domain of group II intron IEPs are important for mediating interactions with intron RNA that promote mobility [Blocker *et al.* 2005]. Thus, we set out to determine the structures of group II intron IEPs with and without bound intron RNA substrates. High-resolution structures were to be solved by X-ray crystallography, with small angle X-ray scattering (SAXS) experiments providing complementary structural data as well as information about conformational changes that occur upon substrate binding and catalysis. Unfortunately, attempts to crystallize group II intron IEPs and RNP complexes were ultimately unsuccessful. However, SAXS studies in conjunction with homology modeling have provided new insights into the structure and function of group II intron IEPs. SAXS data were used to generate low-resolution shape reconstructions of GsI-IIC, while high-quality homology models were used to propose a mechanism of interaction between the IEP and the RNA-DNA hybrid duplex formed during retrohoming.

Additionally, these structures will serve to increase our knowledge of other non-LTR retroelement RTs, an important class of enzymes for which there are no known structures and whose properties differ significantly from those of retroviral RTs [Lambowitz and Zimmerly 2004 and Lambowitz and Zimmerly 2010].

## Chapter 2: Experimental Design and Methods

### SAMPLE PREPARATION

#### Expression of group II intron-encoded proteins

Historically, active group II intron-encoded reverse transcriptases (RTs) have been difficult to purify free of bound intron RNA [Mohr *et al.* 2013]. Consequently, past attempts to investigate the structure of these important enzymes have been limited in scope. Therefore, it was necessary to develop general methods for the large-scale production and purification of group II intron RTs. It had been shown previously that the expression and solubility of certain proteins could be improved by fusion of a highly soluble protein, like maltose-binding protein (MBP) [Nallamstetty and Waugh 2006]. In addition to enhancing protein solubility, MBP tags allow for efficient protein purification via amylose-affinity chromatography.

To test whether group II intron RTs could be expressed and purified as MBP fusions, Mohr *et al.* developed a protocol that includes polyethylenimine (PEI) precipitation, amylose-affinity chromatography, and a heparin-Sepharose purification step [Mohr *et al.* 2013]. The PEI-precipitation step removes bound nucleic acids that contribute to sample heterogeneity. Preliminary experiments in which the MBP tag was fused to the N-terminus of the RT via a tobacco etch virus (TEV) protease-cleavable linker allowed for efficient expression and purification of active RTs (Figure 7). However, when the MBP tag was removed by protease cleavage, the RTs immediately precipitated and degraded readily. This result was unexpected, as proteins generally retain their solubility after tag removal [Nallamstetty and Waugh 2006], and suggests that group II intron RTs are normally co-expressed with the intron RNA from which they are translated, forming an ribonucleoprotein (RNP) complex.

To resolve these issues, Mohr *et al.* tested whether group II intron RTs could be stabilized by fusing the MBP tag via a non-cleavable, rigid linker [Mohr *et al.* 2013]. These types of rigid fusions are traditionally used to enhance conformational homogeneity for protein crystallization, and they generally contain a short 5-alanine linker (Figure 7) [Smyth *et al.* 2003]. Mohr *et al.* found that thermostable group II intron RTs were readily expressed as MBP rigid fusions and exhibited near wild-type efficiency in retrohoming assays, suggesting that they retain all required activities, despite the presence of the rigidly fused tag [Mohr *et al.* 2013].

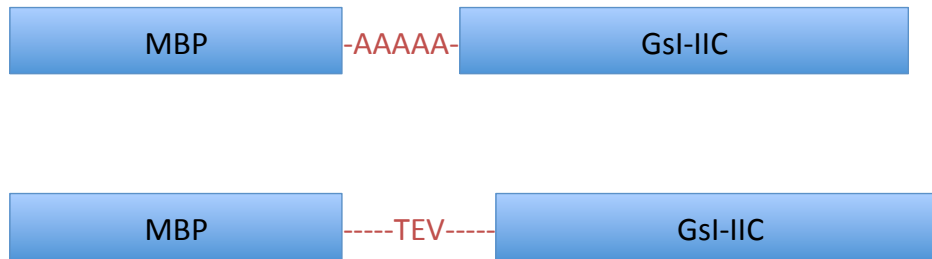


Figure 7: Schematic of MBP-GsI-IIC fusion proteins. The maltose rigid fusion MRF-GsI-IIC construct (top) has an N-terminal maltose binding protein (MBP) tag that is rigidly fused via a short, non-cleavable 5-alanine linker. The maltose fusion MBP-GsI-IIC (bottom) has an N-terminal MBP tag that is fused via a flexible tobacco etch virus (TEV) protease-cleavable linker.

### Cloning of expression plasmids

The pMalE-GsI-IIC construct contains the GsI-IIC open reading frame (ORF) (from Mohr *et al.*) with an N-terminal MBP tag cloned behind the *tac* promoter in pMal-c2t (derived from pMal-c2x; New England BioLabs) [Mohr *et al.* 2013 and Kristelly *et al.* 2003]. pMalE-GsI-IIC was constructed by PCR amplifying the ORF from *Geobacillus stearothermophilus* strain 10 genomic DNA (obtained from Greg Davis, Sigma-Aldrich) and cloning the PCR products into pMal-c2t. GsI-IIC is a group IIC intron found in multiple copies in the *G. stearothermophilus* genome (Moretz and Lampson 2010). The

cloned GsI-IIC ORF corresponds to one of these genomic sequences and has three amino acid sequence changes compared with a related RT ORF cloned by Vellore *et al.* [Vellore *et al.* 2004].

The pMRF-GsI-IIC construct contains the GsI-IIC ORF with an MBP tag linked in frame with the N terminus of the ORF via a rigid fusion. It was derived from the corresponding pMalE plasmids by replacing the TEV protease-cleavable linker (TVDEALKDAQTNS<sub>3</sub>N<sub>10</sub>LENLYFQG) with a rigid linker (TVDAALAAQTAAAAA) by using QuikChange PCR mutagenesis with Accuprime polymerase (Life Technologies) [Mohr *et al.* 2013 and Makarova *et al.* 2000].

### **Protein expression and purification**

Group II intron-encoded fusion proteins MBP-GsI-IIC and MRF-GsI-IIC were expressed in *E. coli* Rosetta 2 (EMD Chemicals). *E. coli* were transformed with the expression plasmid and grown at 37°C in 500-mL TB medium in 2.5-L Ultrayield flasks (Thompson Instrument Company) or 1-L LB medium in 4-L Erlenmeyer flasks. Expression was induced by adding isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG; 1 mM final) to mid-log phase cells ( $OD_{600} = 0.8$ ). Cells were induced at 22°C for 5 h, pelleted by centrifugation, resuspended in buffer A (20 mM Tris-HCl [pH 7.5], 0.5 M KCl, 1 mM EDTA, 1 mM dithiothreitol [DTT]), and frozen at -80°C.

Fusion proteins were purified by a procedure that involves cell disruption by freeze-thawing and sonication; polyethylenimine (PEI) precipitation of nucleic acids; amylose-affinity chromatography; and heparin-Sepharose chromatography. The cell suspension was thawed, treated with lysozyme (1mg/mL; Sigma) for 30 min on ice, then subjected to three cycles of freeze-thawing on dry ice, followed by sonication (Branson 450 Sonifier, amplitude 60% on ice; six 15-sec burst with 10-sec in between bursts).

After centrifugation to pellet cell debris, nucleic acids were precipitated by adding PEI to a final concentration of 0.4% and centrifugation at 15,000g for 15 min at 4°C (J16.25 rotor; Avanti J-E centrifuge; Beckman Coulter). The resulting supernatant was loaded onto an amylose column (Amylose High-Flow; New England BioLabs; 12-mL column equilibrated in buffer A, which was then washed with five column volumes each of buffer A containing 0.5 M KCl, and eluted with buffer A containing 10 mM maltose. Protein fractions were pooled and purified further by heparin-Sepharose chromatography (5-mL column; GE Healthcare Biosciences). The heparin-Sepharose column was equilibrated and the samples were loaded in 20 mM Tris-HCl (pH 7.5), 500 mM KCl, 1 mM EDTA, 1 mM DTT, 10% glycerol. The proteins were applied to the column in the same buffer and eluted with 40-column volume KCl gradient from the loading concentration of 2 M. Peak fractions of the RTs, which eluted at ~800 mM KCl, were pooled and dialyzed against 20 mM Tris-HCl [pH 7.5], 0.5 M KCl, 1 mM EDTA, 1 mM DTT, and 10% glycerol. The proteins were further purified to homogeneity by gel filtration (GE Healthcare) or flash frozen in liquid N<sub>2</sub> and stored at -80°C.

Protein concentrations were determined either by using Bradford assay [Bradford 1976] with bovine serum albumin (BSA) as a standard or by using the Qubit fluorescent assay according to manufacturer's instructions (Life Technologies). A unit of RT activity is defined by the amount of enzyme required to polymerize 1 nmol of dTTP in 1 min at 60°C, using poly(rA)/olig(dT)<sub>42</sub> as template, as described by Mohr *et al.* [Mohr 2013]. All protein preparations were >95% pure, and the yields of MBP-GsI-IIC grown in TB medium in Ultrayield flasks were 5-10 mg/L.

## **Group II intron ribonucleoprotein preparation**

Purified group II intron precursor RNAs were incubated with purified IEPs at various temperatures in Buffer A to promote splicing and *in vitro* assembly of ribonucleoproteins (RNPs) for crystallization and small angle X-ray scattering experiments. The RNPs were further purified to homogeneity by gel filtration (GE Healthcare) and immediately frozen with liquid N<sub>2</sub> and stored at -80°C for SAXS experiments or stored at 4°C for crystallization trials. The *G. stearothermophilus* group II introns used for RNP complex formation lack the 1414 nucleotide ORF of Domain IV (DIV), and this construct is described in Figure 8.

## ***In vitro* synthesis of group II intron precursor RNA**

Precursor RNAs containing group II introns (Figure 8) were synthesized using *in vitro* run-off transcription by phage T7 RNA polymerase and purified by gel filtration chromatography as described by McKenna and colleagues [McKenna *et al.* 2007]. For this procedure, the template plasmid pUC19-GsI-IIC was transformed into *E. coli* DH5 $\alpha$  competent cells (Life Technologies) and a single colony was used to inoculate a 50-mL LB culture that was grown for 12 hours at 37°C. The 50-mL culture was diluted into 3 L of fresh LB medium and grown for 18 h at 37°C in a 4 L flask. The template plasmid was purified using a commercially available kit according to manufacturer's instructions (QIAGEN) and were stored at -20°C. The template was linearized by exhaustive digest with BamHI, and complete linearization was confirmed by 1% agarose gel electrophoresis. 3 M sodium acetate [pH 5.2] was added to the digestion reaction to a final concentration of 300 mM. Plasmid DNA was precipitated by adding a fourfold volume of cold ethanol. Linearized plasmids were pelleted by centrifugation at 35,000g for 30 min at 4°C. The supernatant was discarded and the pellet retained and washed with 10 mL of cold 70% ethanol. Excess liquid was removed and the pellet was allowed to dry

before being resuspended in 1 mL of deionized water. It was then diluted to 500 µg/mL and stored at -20°C.

Small-scale (50 µL) transcription reactions were prepared in a 1X transcription buffer (400 mM Tris-HCl pH 8.1, 10 mM spermidine, 0.01% (wt/vol) Triton X-100 and 100 mM DTT), 8 mM NTPs (ATP, GTP, CTP, UTP; 2 mM each) and 2.5 µg of linearized template brought to a final volume of 50 µL in deionized water. T7 RNA (produced in house) polymerase was included at 0.5, 1.0, or 2.0 µL per 50 µL reaction. For increasing amounts of T7 RNA polymerase, the concentration of MgCl<sub>2</sub> was adjusted incrementally from 5 to 50 mM. Trial transcription was performed at 37° C for 1 h. To assess transcription levels, 10 µL aliquots were assayed by denaturing polyacrylamide gel electrophoresis. The gel was stained with 0.1% toluidine blue solution for 5 min and then destained. Transcripts synthesized from the template that had been completely linearized ran as a single band.

This reaction was scaled up to 10 mL (in a 50 mL conical vial) and transcription was performed for 1 h at 37°C. Pyrophosphate was removed by centrifugation at 3000g for five minutes at room temperature. The supernatant was retained and EDTA was added to a final concentration of 50 mM to chelate magnesium. An equal volume of phenol/chloroform 1:1 was added to quench the reaction, and the tube was inverted and centrifuged at 3000g for 10 min at room temperature. The aqueous phase was retained thus removing T7 RNA polymerase and BamHI. The extraction was repeated two more times to ensure efficient removal of contaminating enzymes.

Transcripts were desalted on a 10-DG column (BioRad) equilibrated in a buffer containing 10 mM sodium phosphate pH 6.6 and 100 mM KCl (RNA buffer). 3 mL of the aqueous phase was loaded onto the column, and the column was drained. 1 mL of the RNA buffer was loaded and drained. The RNA was then eluted with 5 mL of RNA buffer



until all of the aqueous phase was consumed. The desalted transcripts were then loaded onto a size exclusion column (GE Healthcare) to separate the RNA from template DNA and any abortive transcripts. Purity was assayed by native gel electrophoresis and RNA concentrations were measured using a NanoDrop 1000 according to manufacturer's instructions (Thermo Scientific). Purified RNAs were then stored at 4°C.

### **Cloning of RNA template plasmids**

The pUC19-GsI-IIC construct contains the GsI-IIC intron sequence, lacking the ORF encoded within DIV, and flanked by shorter than wild type (WT) exons, cloned behind a T7 promoter in pUC19 (New England BioLabs) with primers that append a BamHI restriction site to the end of the 5' exon. This construct was used as a DNA template for *in vitro* transcription of group II intron precursor RNA. The pUC19-GsI-IIC plasmid was constructed by PCR amplifying the appropriate sequences from *G. stearothermophilus* strain 10 genomic DNA (obtained from Greg Davis, Sigma-Aldrich) and cloning the PCR products into pUC19 (New England BioLabs). This construct contains shorter than WT exons, as these lengths (35 and 45 nts for the 5' and 3' exons, respectively) have been shown to promote efficient splicing *in vitro* [Qin *et al.*, unpublished work].

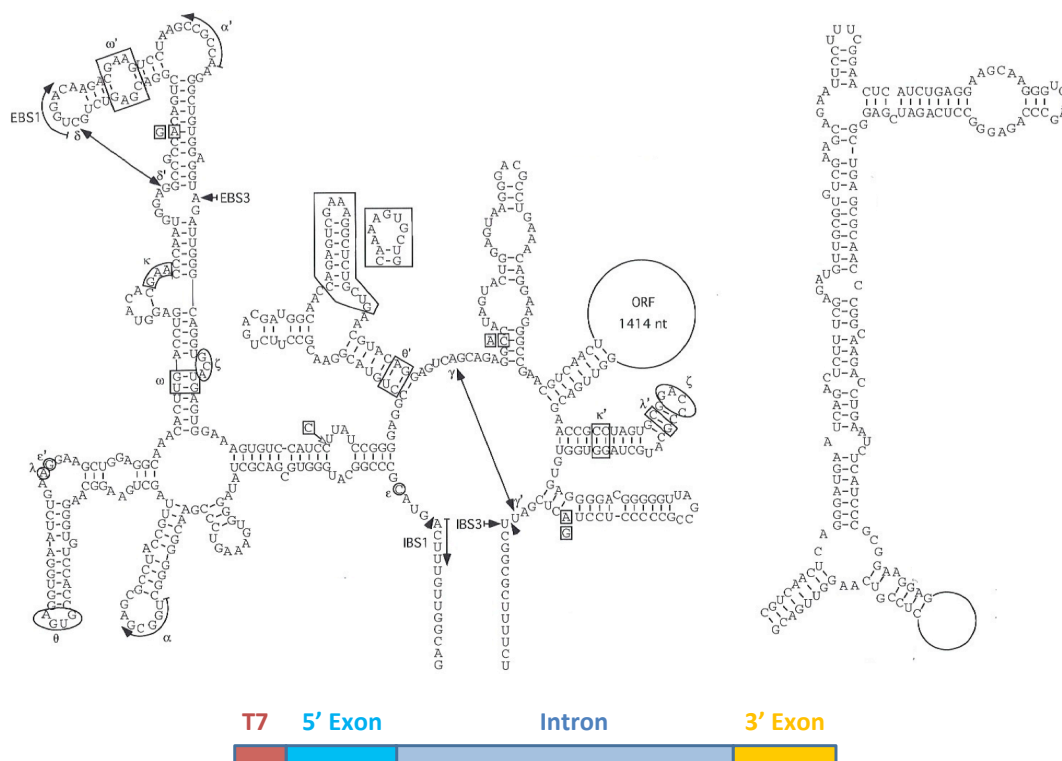


Figure 8: Model of the GsI-IIC intron and schematic representation of the precursor RNA construct used for RNP complex formation. The predicted secondary structure of the GsI-IIC intron (top left) contains six double-helical domains. The RNA construct used for RNP complex formation (bottom) lacks the GsI-IIC ORF in DIV. The remaining portion of DIV is shown in the top right.

### Synthetic substrate mimics

Synthetic DNA/RNA heteroduplexes (Figure 9) designed to promote RT active site closing served as artificial RT substrates for crystallization trials and SAXS experiments. Methods for forming closed RT complexes are described in Figure 9. To assemble the heteroduplexes, synthetic oligonucleotides (Integrated DNA Technologies) were re-suspended in sterile TE buffer (QIAGEN) to 2 mM and annealed by heating to 95°C for 1 minute and cooling to room temperature over 1 h in a thermal cycler (Eppendorf). Freshly prepared MRF-GsI-IIC (~150  $\mu$ M) was incubated with a molar excess of annealed heteroduplex, 2 mM  $MgCl_2$ , and 2 mM nucleotides for 10 min at

25°C, spin filtered (Millipore), and purified by gel filtration (GE Healthcare). Individual fractions were frozen with liquid N<sub>2</sub> and stored at -80°C for SAXS. For crystallography, the cleanest fractions were pooled, concentrated to ~15 mg/ml, and stored at 4°C for crystallography. Samples for SAXS experiments were further concentrated just before data collection to obtain buffer blanks and check for any concentration-dependent scattering effects.

#### 1. Use dideoxy nucleotide in DNA strand



#### 2. Let RT add 2 Gs and then stop elongation with ddC



#### Alternate RNA strand lengths



Figure 9: Synthetic substrates mimics. Hybrid duplexes used for synthetic complex formation consist of one ssDNA strand (black) annealed to either a 17, 25, 35, or 45 nt RNA strand (red). The DNA strand in (1) contains a 3' dideoxy nucleotide and should bind along with dGTP at the active site of GsI-IIC to promote active site closing as the RT stalls in an attempt to add nucleotides to the 3' end of the DNA strand. The DNA strand in (2) lacks a 3' dideoxy nucleotide and should bind the active site of GsI-IIC, which can add two dG nucleotides to the 3' end of the DNA strand and then stall in a closed conformation after the incorporation of ddC.

### X-RAY CRYSTALLOGRAPHY

The full-length GsI-IIC crystallography construct was fused to MBP to allow for streamlined affinity column chromatography purification and to maintain sample solubility [Smyth 2003]. The MBP tag was retained for crystallization experiments as a

method for promoting crystal growth, as large affinity tags had previously been shown to help direct the formation of crystal contacts in fusion proteins [Smyth *et al.* 2003 and Moon *et al.* 2010]. In these constructs, the IEP was rigidly fused to MBP via a non-cleavable linker that is shorter than the original linker in the pMAL-c2x vector from New England BioLabs. This modification was made to maintain sample rigidity for crystallization, and these MBP tags have also been engineered such that their overall surface entropy is reduced relative to the WT protein [Moon *et al.* 2010]. These tags have been used to solve the structures of three unrelated proteins [Smyth *et al.* 2003].

Purified MRF-GsI-IIC fusions and RNP complexes dialyzed into a buffer containing 20 mM Tris-HCl [pH 7.5], 250 or 500 mM KCl, and 10% glycerol were to be crystallized by common vapor diffusion methods as previously described [Del Campo *et al.* 2009]. Briefly, initial screening was done by hanging-drop methods in 96-well plates at 22°C with commercially available kits from Hampton Research (Crystal Screen HT, Index HT, PEG/ION HT, Salt Rx HT, and PEG Rx HT). For each screen, the concentration of protein was varied from 5 to 20 mg/mL and the reservoir contained 50  $\mu$ L of the screening buffer. Optimization of crystallization conditions was to be performed by sitting-drop methods, and improved single crystals were to be flash frozen with liquid N<sub>2</sub>. X-ray diffraction data was to be collected at a synchrotron source.

### **Surface entropy reduction**

In an effort to increase the probability of crystallizing the MRF-GsI-IIC fusion protein, several alternative constructs were engineered to reduce the conformational entropy of GsI-IIC. In these constructs, several large hydrophilic surface residues are mutated to alanine, a small nonpolar amino acid. These mutations enhance the likelihood of crystal lattice formation, by reducing conformational entropy at the protein surface

[Goldschmidt *et al.* 2007]. To identify suitable sites for mutating specific surface residues in GsI-IIC, the ORF was submitted to the online surface entropy prediction (SERp) server developed by Goldschmidt *et al.* The server assigns a score to each residue in the sequence and identifies those most favorable for mutation [Goldschmidt *et al.* 2007]. These results were used to engineer a set of five alanine mutant constructs (Table 2). These pSER-GsI-IIC constructs were derived from the pMRF-GsI-IIC plasmid by mutating selected residue(s) to alanine using QuikChange PCR mutagenesis according to manufacturer's instructions (Agilent Technologies). The mutagenic primers used for this procedure are summarized in Table 1.

Mutations	Primers
E186A E187A	Forward: 5'-AAGGGGTGAAGGTGCAGACGGCTGCTGGGACGCCGCAAGGCCG-3' Reverse: 5'-CCGCCTTGCGGCGTCCCAGCAGCCGTCTGCACCTTCACCCCTT-3'
E212A K213A	Forward: 5'-GATTTAGACAAGGAATTGGCCGCGCGAGGATTGAAATTCTGC-3' Reverse: 5'-GCAGAATTTCAATCCTCGCGCGGCAATTCCTTGTCTAAATC-3'
E256A E257A	Forward: 5'-CAAACCTCAAAGTAAACGCGCGGCGAAAAGTGCGGTGGACCG-3' Reverse: 5'-CGGTCCACCGCACTTTTCGCGCGGTTTACTTTGAGTTTG-3'
K258A	Forward: 5'-AAAGTAAACGAGGAGGCTAGTGCGGTGGACCGC-3' Reverse: 5'-GCGGTCCACCGCACTAGCCTCCTCGTTTACTTT-3'

Table 1: Mutagenic primers used for generating pSER-GsI-IIC constructs.

## SMALL ANGLE X-RAY SCATTERING

In small angle X-ray scattering experiments, monochromatic X-rays are elastically scattered by a solution of macromolecules (Figure 10A) [Lipfert and Doniach 2007]. The observed scattering intensity is the sum of the scattering intensity from the randomly oriented molecules that populate the solution [Lipfert and Doniach 2007]. Unlike scattering by a crystal, there is no constructive interference of X-rays scattered by adjacent molecules [Lipfert and Doniach 2007]. The intensities of the scattered rays (I)

are recorded as a function of the scattering angle ( $2\theta$ ), and the resulting scattering profile contains information on the size and shape of the molecules in solution (Figure 10).

Recently, I collected synchrotron radiation X-ray scattering data from MRF-TeI4c and MRF-GsI-IIC at the Advanced Light Source SIBYLS beamline and the Advanced Photon Source 12-ID-C beamline as previously described [Mallam *et al.* 2011]. The data were recorded with a two-dimensional charge coupled device detector. Twenty separate one-second exposures were acquired for each sample at a sample-to-detector distance of  $\sim 2.0$  m over a range of momentum transfer  $\sim 0.005 < q < \sim 0.30 \text{ \AA}^{-1}$ , where  $q = 4\pi\sin(\theta)/\lambda$ , and  $\lambda$  is the X-ray wavelength. Scattering data were radially averaged to produce one-dimensional profiles of scattering intensity ( $I$ ) versus momentum transfer ( $q$ ). Data were collected for buffer blanks and increasing protein concentrations to check for concentration-dependent scattering effects such as aggregation and inter-particle interference. The zero-angle scattering intensity ( $I(0)$ ) was calibrated against known concentrations of bovine serum albumin standards.

Figure 10B shows typical background-subtracted scattering intensities plotted as a function of momentum transfer in IGOR-Pro (WaveMetrics). Information about particle size and sample homogeneity can be extracted from the low-resolution data [Lipfert and Doniach 2007]. Linearity in the Guinier plot ( $\ln[I(q)]$  versus  $q^2$ ) at low  $q$  values indicates a monodisperse sample free from aggregation (Figure 10C) [Jacques and Trewhella 2010]. Using the ATSAS software suite (version 2.4), the radius of gyration ( $R_g$ ) can be calculated from low-resolution data using the Guinier approximation [Jacques and Trewhella 2010]. The radius of gyration is defined as the root mean square average of the distance of the scattering elements within the particle from the center of the particle, and this value provides an estimate of particle size [Lipfert and Doniach 2007]. To properly estimate the radius of gyration,  $R_g$  must be calculated as a function of concentration and

then extrapolated to infinite dilution to ensure that there is no inter-particle scattering [Lipfert and Doniach 2007].

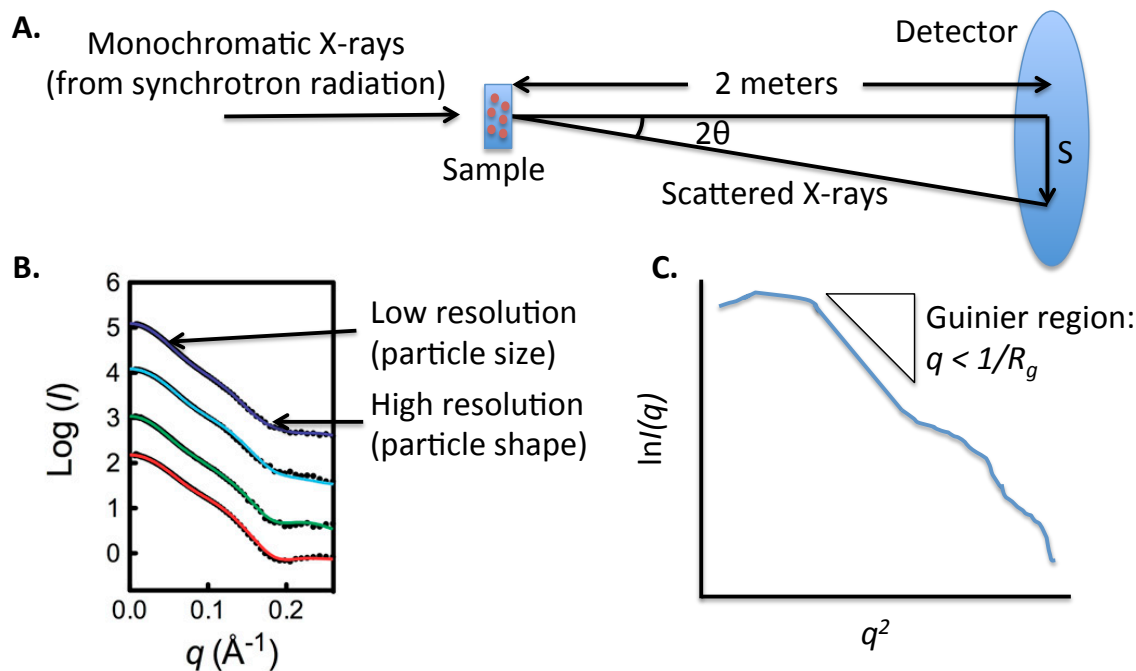


Figure 10: SAXS experimental set up and data collection. (A) In a SAXS experiment, monochromatic X-rays are scattered by a solution of macromolecules. (B) Scattering intensity is plotted as a function of the momentum energy transfer  $q = 4\pi\sin(\theta)/\lambda$ , where  $2\theta$  is the scattering angle and  $\lambda$  is the X-ray wavelength. Adapted from Mallam *et al.* 2010. (C) In a Guinier plot, the natural log of the scattering intensity is plotted as a function of  $q^2$ .

From the scattering curve it is possible to directly calculate a distance distribution function  $P(r)$  (Figure 11). This function is a Fourier transform of the scattering data and describes the probability of finding an electron separated by a distance  $r$  from another electron in the object [Lipfert and Doniach 2007]. The hypothetical distance distribution functions for a spherical particle of 30 Å and a particle containing two spherical domains of 30 Å each separated by a long linker are shown in Figure 11B. These figures serve to illustrate the meaning of the distance distribution function and how information on

particle shape can be extracted from SAXS data. The distance distribution function in Figure 11A was calculated with the GNMO algorithm from Semenyuk and Svergun using the program AUTOGNOM [Jacques and Trewhella 2010 and Petoukhov *et al.* 2007]. Implementing this program also provides the maximum particle dimension  $D_{\max}$  and a value for  $R_g$  calculated from the whole scattering profile (real-space  $R_g$ ).

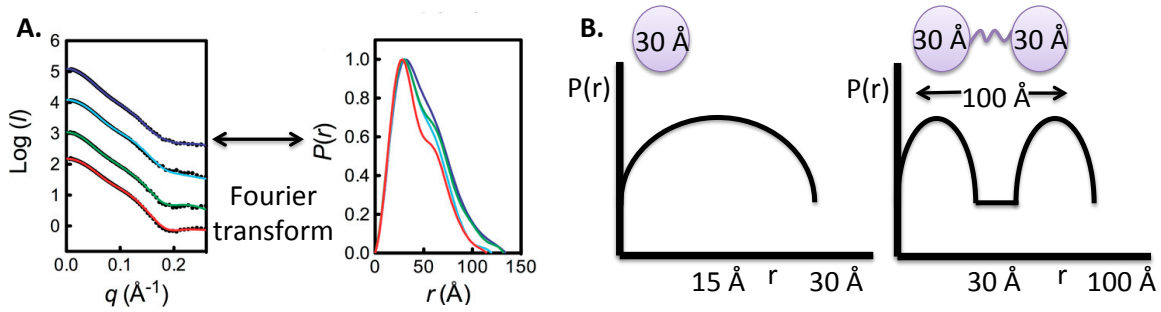


Figure 11: The distance distribution function. (A) The distance distribution function is a Fourier transform of the scattering curve. Adapted from Mallam *et al.* 2011. (B) Two hypothetical distance distribution functions illustrate how information on particle shape can be extracted from SAXS data.

It is also possible to generate three-dimensional reconstructions of the SAXS data using *ab initio* modeling methods. These low-resolution envelopes are constructed by simulated annealing procedures using the programs DAMMIN and GASBOR [Svergun 1999 and Svergun and Petoukhov 2001]. The DAMMIN procedures start with a dense, spherical package of beads that constitute an initial search volume for generating a dummy atom protein model [Svergun 1999]. GASBOR uses a chain-like assembly of dummy residues for shape reconstruction. This method ensures that the reconstructed model maintains peptide-like properties [Svergun and Petoukhov 2001]. These analyses must be run several times independently until a reproducible result is achieved. The constructions are then averaged using DAMAVER [Volkov and Svergun 2003]. This procedure provides a quantitative measure of the similarity between reconstructions of



the same resolution. The averaged reconstruction can then be refined using DAMMIN or GASBOR to produce a low-resolution model that fits the experimental data [Volkov and Svergun 2003].

## **MOLECULAR MODELING**

Ultimately, we would like to have solved the crystal structure of a group II intron-encoded RT. In lieu of a crystal structure, the best we can do is to develop a structural model of group II intron RTs. Computational methods for predicting three-dimensional protein structures are generally divided into three categories, comparative modeling, threading methods, and *ab initio* modeling [Roy *et al.* 2010]. Alternatively, there are composite approaches that allow for better modeling. The iterative threading assembly refinement (I-TASSER) method is one such composite approach to structure modeling and has been consistently ranked as the best method for high-quality, automated protein structure prediction [Zhang 2008 and Roy *et al.* 2010]. The I-TASSER server is an online platform for protein structure and function prediction developed by the Zhang Lab at the University of Michigan. When a user submits an amino acid sequence, the I-TASSER server generates a three-dimensional (3D) atomic model based on multiple threading alignments and iterative fragment assembly simulations [Zhang 2008 and Roy *et al.* 2010]. The biological function of the protein is then predicted by structurally matching the model to proteins of known function [Zhang 2008 and Roy *et al.* 2010]. The accuracy of this prediction is determined from the confidence score of the modeling procedure [Zhang 2008 and Roy *et al.* 2010]. The I-TASSER method can be divided into four general steps (Figure 12), and a brief description of each step (adapted from Roy *et al.* 2010) is provided here.

In the first step of the I-TASSER protocol, a user-submitted sequence is matched to a database, in order to identify any evolutionarily related sequences. Next, a sequence profile is generated based on multiple sequence alignment of the identified homologs [Roy *et al.* 2010]. Using this sequence profile, secondary structure predictions are made. Taking this information into account, a locally installed meta-threading server (LOMETS) threads the query sequence through a representative database of known protein structures. This threading procedure identifies template proteins from the Protein Data Bank (PDB) structure library that have similar structural motifs as the query sequence [Roy *et al.* 2010]. LOMETS relies on ten different threading programs, and hits from each individual program are ranked using several sequence-based and structure-based scoring techniques. The top hits from each program are selected for further analysis [Zhang 2008]

In the second step, fragments from the selected threading alignments are isolated from the template structures. These fragments are used to assemble structural conformations for well-aligned sections, while unaligned sections are constructed by *ab initio* modeling [Zhang 2008]. This fragment assembly procedure uses a modified Monte Carlo simulation technique [Roy *et al.* 2010]. In order to identify low free-energy states, the generated conformations are clustered. Then, cluster centroids are calculated by averaging the 3D coordinates of the clustered structures [Zhang 2008].

In the third step, fragment assembly simulations are carried out a second time, beginning with selected cluster centroids. However, these simulations are subject to additional constraints from the PDB templates that are structurally similar to the cluster centroids [Roy *et al.* 2010]. Structurally similarity is determined using TM-align [Zhang 2008]. This second round of simulations serves to remove steric clash. The conformations generated during this iteration are clustered, and the lowest free-energy states are used to

generate final structural models with optimized hydrogen bonding networks [Roy *et al.* 2010].

In the fourth and final step, the function of the query protein is determined by structurally matching the 3D model to proteins of known structure and function from the PDB. To estimate the accuracy of the prediction, a confidence score (C-score) is given that allows users to assess the quality of the final model [Roy *et al.* 2010]. The I-TASSER server output includes full-length secondary and tertiary structure predictions as well as functional annotations.

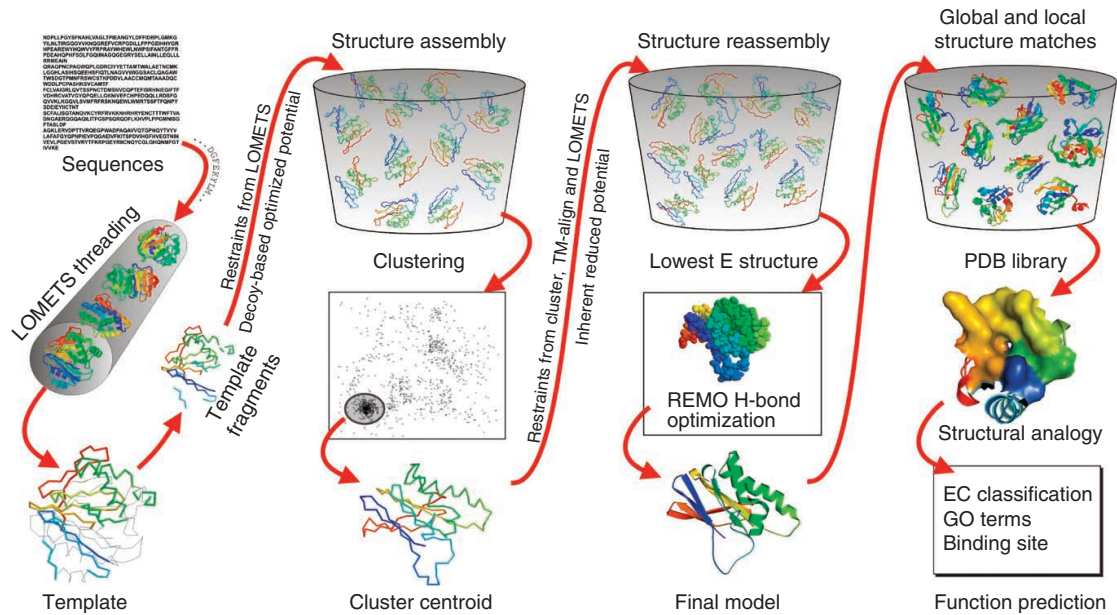


Figure 12: Schematic representation of the I-TASSER protocol. Adapted from Roy *et al.* 2010.

## Chapter 3: Results and Discussion

### SAMPLE PREPARATION

To date, we have successfully purified large quantities of GsI-IIC RT, a group II intron-encoded protein from the thermophilic bacterium *Geobacillus stearothermophilus*. The full length GsI-IIC protein was expressed as an maltose binding protein (MBP) rigid fusion (MRF-GsI-IIC) and purified to homogeneity by affinity chromatography and gel filtration (Figure 13). These methods have significantly improved yields of catalytically active group II IEPs (approximately 6 mg/L of cell culture). Additionally, large quantities of precursor RNAs containing the GsI-IIC group II intron have been transcribed *in vitro* and purified by gel filtration. To inhibit hydrolytic intron self-splicing that occurs during transcription, the concentration of free  $Mg^{2+}$  in these reactions was reduced significantly. This low concentration of  $Mg^{2+}$  sacrifices transcriptional efficiency but ensures sample homogeneity. Unfortunately, the purification of concentrated RNP samples (assembled *in vitro*) in sufficient quantities for crystallography and SAXS experiments has proven difficult, as the resulting complexes are highly insoluble. Presently, we are working to find suitable reaction conditions that promote efficient splicing and maintain sample solubility at protein concentrations greater than 2 mg/mL. Despite this initial set back, we have successfully purified MRF-GsI-IIC bound to synthetic heteroduplexes designed to promote RT active site closing. Crude gel shift assays indicate complex formation, and this result has been further verified by gel filtration (Figure 13).

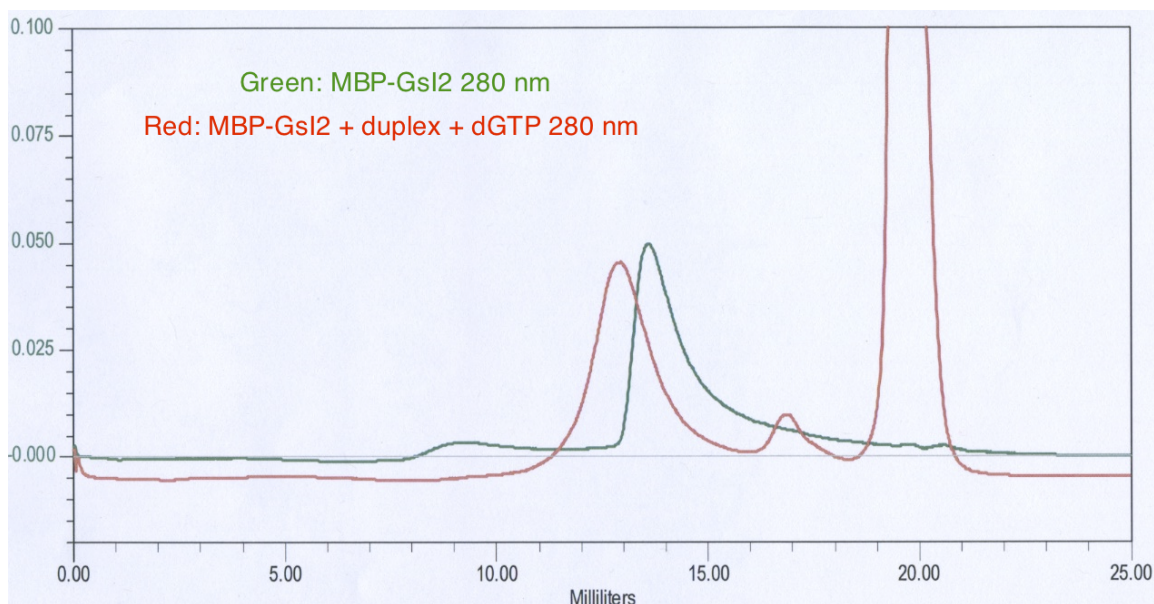


Figure 13: Gel filtration chromatogram. Purified MRF-GsI-IIC elutes from a gel filtration column as a single peak (green curve). Purified MRF-GsI-IIC bound to a synthetic heteroduplex elutes as a single peak, demonstrating complex formation (red curve). The small peak at 17 mL corresponds to excess, unbound heteroduplex and the larger peak at 20 mL corresponds to excess, unbound dGTP.

### X-RAY CRYSTALLOGRAPHY

Crystallization trials with MRF-GsI-IIC with and without bound nucleic acid substrate have not yielded any crystals with desirable qualities for X-ray crystallography. Recently, the online surface entropy reduction prediction (SERp) server was used to identify suitable sites for mutating specific unstructured surface residues in GsI-IIC that likely inhibit the formation of a crystal lattice [Goldschmidt *et al.* 2007]. These results were used to generate a set of alanine mutant constructs with lower conformational entropy for further crystallization trials (Table 2). We are also developing alternative fusion constructs to promote crystallization.

Construct	Sequence
1	MALLERILARDNLITALKRVEANQGAPGIDGVSTDQLRDYIRAHWSTIHAQLLAGTYRPA PVRRVEIPKPGGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSSSSFGFRPGRNAHDAV RQAQGYIQEGYRYVVDMDLEKFFDRVNHDIILMSRVARKVKDKRVLKLIRAYLQAGVMIEG VKV <b>QTEEG</b> TPQGGPLSPLLANILLDDLD <b>KELEK</b> RGLKFCRYADDCNIYVKSLRAGQRVKQ SIQRF <b>EKTLKLKVNEEKSA</b> VDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQIRIQLT NPNWSISMPERIHRVNQYVMGWIGYFRLVETPSVLQTIEGWIRRRRLRCQWLQWKRVRTR IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQALGKTYWTAQGLKSLTQRYFELRQG
2	MALLERILARDNLITALKRVEANQGAPGIDGVSTDQLRDYIRAHWSTIHAQLLAGTYRPA PVRRVEIPKPGGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSSSSFGFRPGRNAHDAV RQAQGYIQEGYRYVVDMDLEKFFDRVNHDIILMSRVARKVKDKRVLKLIRAYLQAGVMIEG VKV <b>QTEEG</b> TPQGGPLSPLLANILLDDLD <b>KELEK</b> RGLKFCRYADDCNIYVKSLRAGQRVKQ SIQRF <b>EKTLKLKVNEEKSA</b> VDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQIRIQLT NPNWSISMPERIHRVNQYVMGWIGYFRLVETPSVLQTIEGWIRRRRLRCQWLQWKRVRTR IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQALGKTYWTAQGLKSLTQRYFELRQG
3	MALLERILARDNLITALKRVEANQGAPGIDGVSTDQLRDYIRAHWSTIHAQLLAGTYRPA PVRRVEIPKPGGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSSSSFGFRPGRNAHDAV RQAQGYIQEGYRYVVDMDLEKFFDRVNHDIILMSRVARKVKDKRVLKLIRAYLQAGVMIEG VKV <b>QTEEG</b> TPQGGPLSPLLANILLDDLD <b>KELEK</b> RGLKFCRYADDCNIYVKSLRAGQRVKQ SIQRF <b>EKTLKLKVNEEKSA</b> VDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQIRIQLT NPNWSISMPERIHRVNQYVMGWIGYFRLVETPSVLQTIEGWIRRRRLRCQWLQWKRVRTR IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQALGKTYWTAQGLKSLTQRYFELRQG
4	MALLERILARDNLITALKRVEANQGAPGIDGVSTDQLRDYIRAHWSTIHAQLLAGTYRPA PVRRVEIPKPGGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSSSSFGFRPGRNAHDAV RQAQGYIQEGYRYVVDMDLEKFFDRVNHDIILMSRVARKVKDKRVLKLIRAYLQAGVMIEG VKV <b>QTEEG</b> TPQGGPLSPLLANILLDDLD <b>KELEK</b> RGLKFCRYADDCNIYVKSLRAGQRVKQ SIQRF <b>EKTLKLKVNEEKSA</b> VDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQIRIQLT NPNWSISMPERIHRVNQYVMGWIGYFRLVETPSVLQTIEGWIRRRRLRCQWLQWKRVRTR IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQALGKTYWTAQGLKSLTQRYFELRQG
5	MALLERILARDNLITALKRVEANQGAPGIDGVSTDQLRDYIRAHWSTIHAQLLAGTYRPA PVRRVEIPKPGGGTRQLGIPTVVDRLIQQAILQELTPIFDPDFSSSSFGFRPGRNAHDAV RQAQGYIQEGYRYVVDMDLEKFFDRVNHDIILMSRVARKVKDKRVLKLIRAYLQAGVMIEG VKV <b>QTEEG</b> TPQGGPLSPLLANILLDDLD <b>KELEK</b> RGLKFCRYADDCNIYVKSLRAGQRVKQ SIQRF <b>EKTLKLKVNEEKSA</b> VDRPWKRAFLGFSFTPERKARIRLAPRSIQRLKQIRIQLT NPNWSISMPERIHRVNQYVMGWIGYFRLVETPSVLQTIEGWIRRRRLRCQWLQWKRVRTR IRELRALGLKETAVMEIANTRKGAWRTTKTPQLHQALGKTYWTAQGLKSLTQRYFELRQG

Table 2: GsI-IIC SER constructs. The online SERp server was used to identify suitable sites (bold text) for mutating unstructured surface residues in GsI-IIC that may interfere with the formation of a crystal lattice. In the GsI-IIC SER constructs, residues colored red are mutated to alanine in order to lower conformational entropy and promote the formation of a crystal lattice.

## SMALL ANGLE X-RAY SCATTERING

Preliminary data from light scattering experiment indicate that the MRF-GsI-IIC rigid fusion samples exhibit monodispersity and are free of aggregates. Initial inspection of SAXS scattering curves reconfirms sample monodispersity. From the background subtracted curves (Figure 14), the zero-angle scattering intensity  $I(0)$  and the radius of gyration  $R_g$  were calculated for MRF-GsI-IIC using the Guinier approximation [Jacques and Trewhella 2010]. The Guinier plots exhibit linearity and further confirm sample monodispersity. The molecular weights were estimated from  $I(0)$  and are in good agreement with values calculated from the amino acid sequence and as estimated by size exclusion chromatography (hydrodynamic radius,  $R_H$ ). The data were then analyzed with the program AUTOGNOM to obtain the distance distribution function  $P(r)$ , a maximum particle dimension  $D_{max}$ , and a real-space value for  $R_g$  [Petoukhov *et al.* 2007]. *Ab initio* reconstructions of the three-dimensional shapes were calculated using two different computational methods (DAMMIN and GASBOR), and the independently reconstructed envelopes are in good agreement [Svergun 1999 and Svergun and Petoukhov 2001]. The low-resolution solution structure generated using the program GASBOR is shown in Figure 14.

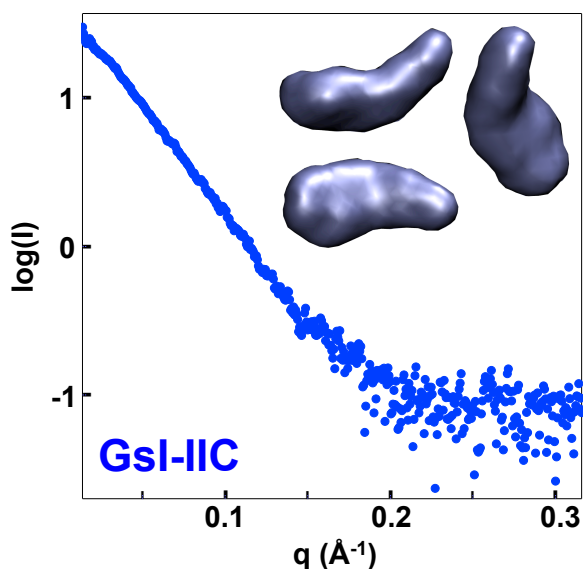


Figure 14: MRF-GsI-IIC SAXS scattering profile and GASBOR shape reconstruction.

Sample	MW (sequence) [kDa]	Abs I(0)/c [cm <sup>2</sup> mg <sup>-1</sup> ]	MW (SAXS) [kDa]	R <sub>g</sub> (Guinier) [Å]	R <sub>g</sub> (GNOM) [Å]	D <sub>max</sub> [Å]
BSA	66	0.0026	----	33.8	----	----
MRF-GsI-IIC	90	0.0025	88	44.2	44.3	142.7

Table 3: MRF-GsI-IIC structural parameters calculated from SAXS data. The molecular weight (MW) of GsI-IIC based on its amino acid sequences is in good agreement with the MW derived from SAXS data. The radius of gyration (R<sub>g</sub>) derived in the Guinier analysis is in good agreement with the real-space value calculated using AUTOGNOM. D<sub>max</sub> is the maximum particle dimension obtained from the distance distribution function.

Sample	R <sub>H</sub> [Å]	R <sub>g</sub> [Å]
MRF-GsI-IIC	40.2	39.8

Table 4: Hydrodynamic and gyration radii. The estimated size of MRF-GsI-IIC as determined by gel filtration chromatograph (R<sub>H</sub>) is in good agreement with the radius of gyration (R<sub>g</sub>) derived from SAXS data.



## MOLECULAR MODELING

The iterative threading assembly refinement (I-TASSER) server was used to generate a high-quality, three-dimensional (3D) model of GsI-IIC based on its amino acid sequence (Figure 15). The 3D model was constructed from multiple threading alignments and iterative fragment assembly simulations [Roy *et al.* 2010]. In this model, the RT and X/thumb domains of GsI-IIC form a central cavity, with the short, C-terminal DNA-binding domain attached via a flexible linker. In order to suggest possible orientations of maltose binding protein (MBP) and GsI-IIC in the MRF-GsI-IIC construct, the I-TASSER model of GsI-IIC and the crystal structure of MBP were positioned within the GASBOR shape reconstruction generated from the SAXS scattering profile (Figure 16).

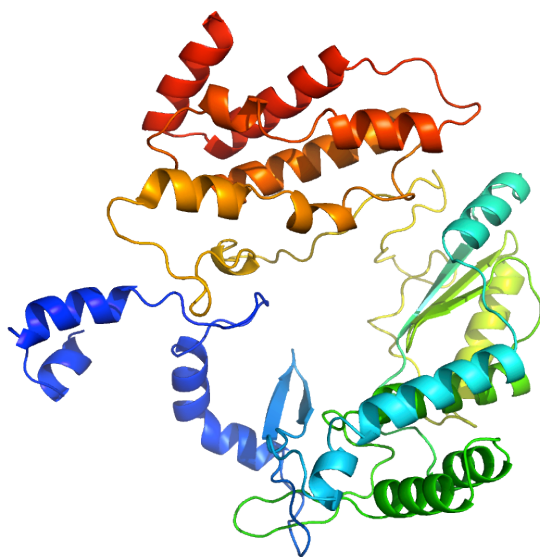


Figure 15: Homology model of GsI-IIC generated by the I-TASSER server.

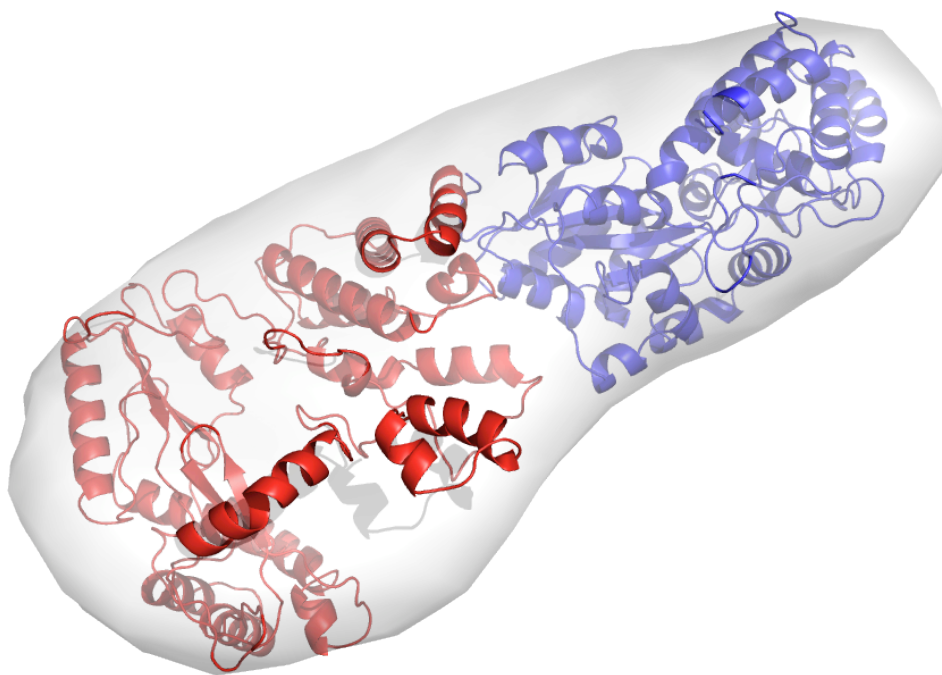


Figure 16: Predicted orientation of maltose binding protein (blue) and full length GsI-IIC (red) positioned within the GASBOR *ab initio* shape reconstruction of MRF-GsI-IIC.

The primary template used in modeling simulations was the crystal structure of the active *Tribolium castaneum* telomerase catalytic subunit, TERT, bound to an RNA-DNA hairpin designed to resemble the TERT substrate (Figure 17). Like group II intron IEPs, telomerase is a specialized RNA-directed DNA polymerase, or reverse transcriptase. In the TERT structure, the RNA-DNA hybrid adopts a helical structure and is docked in the interior cavity of the TERT ring [Mitchell *et al.* 2010]. Contacts between the RNA template and the enzyme position the solvent-accessible RNA bases close to the enzyme active site for nucleotide binding and selectivity [Mitchell *et al.* 2010]. These associations between TERT and its hybrid nucleic acid substrate suggest a possible mechanism of interaction between group II intron IEPs and their RNA-DNA hybrid duplex substrates. It is likely that during the initiation of reverse transcription in

retrohoming, the integrated intron RNA binds to the interior cavity of the IEP. These interactions are strikingly similar to those observed for retroviral RTs [Mitchell *et al.* 2010], suggesting a common mechanism of reverse transcription between all three enzyme families.

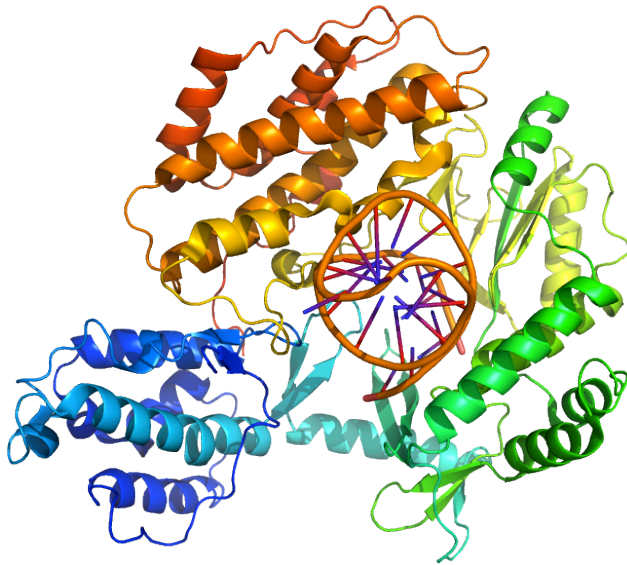


Figure 17: *T. castaneum* telomerase catalytic subunit, TERT bound to a hybrid RNA-DNA duplex substrate mimic (PDB accession 3KYL) [Mitchell *et al.* 2010].

## References

- Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA*. 2005 Jan;11(1):14-28.
- Cavalier-Smith T. Intron phylogeny: a new hypothesis. *Trends Genet*. 1991 May;7(5):145-8.
- Cech TR. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell*. 1986 Jan 31;44(2):207-10.
- Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM. A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. *J Mol Biol*. 2004 Jul 2;340(2):211-31.
- Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ, Lambowitz AM, Zimmerly S. A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. *Mol Cell*. 2008 May 23;30(4):472-85.
- Del Campo M, Lambowitz AM. Crystallization and preliminary X-ray diffraction of the DEAD-box protein Mss116p complexed with an RNA oligonucleotide and AMP-PNP. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2009 Aug 1;65(Pt 8):832-5.
- Goldschmidt L, Cooper DR, Derewenda ZS, Eisenberg D. Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Sci*. 2007 Aug;16(8):1569-76.
- Guo H, Karberg M, Long M, Jones JP 3rd, Sullenger B, Lambowitz AM. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science*. 2000 Jul 21;289(5478):452-7.
- Jacques DA, Trewhella J. Small-angle scattering for structural biology--expanding the frontier while avoiding the pitfalls. *Protein Sci*. 2010 Apr;19(4):642-57.
- Keating KS, Toor N, Perlman PS, Pyle AM. A structural analysis of the group II intron active site and implications for the spliceosome. *RNA*. 2010 Jan;16(1):1-9.
- Kristelly R, Earnest BT, Krishnamoorthy L, Tesmer JJ. Preliminary structure analysis of the DH/PH domains of leukemia-associated RhoGEF. *Acta Crystallogr D Biol Crystallogr*. 2003 Oct;59(Pt 10):1859-62.
- Lambowitz AM, Belfort M. Introns as mobile genetic elements. *Annu Rev Biochem*. 1993;62:587-622.
- Lambowitz AM, Zimmerly S. Mobile group II introns. *Annu Rev Genet*. 2004;38:1-35.
- Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol*. 2011 Aug 1;3(8):a003616.

- Lipfert J, Doniach S. Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annu Rev Biophys Biomol Struct.* 2007;36:307-27.
- Makarova O, Kamberov E, Margolis B. Generation of deletion and point mutations with one primer in a single cloning step. *Biotechniques.* 2000 Nov;29(5):970-2.
- Mallam AL, Jarmoskaite I, Tijerina P, Del Campo M, Seifert S, Guo L, Russell R, Lambowitz AM. Solution structures of DEAD-box RNA chaperones reveal conformational changes and nucleic acid tethering by a basic tail. *Proc Natl Acad Sci U S A.* 2011 Jul 26;108(30):12254-9.
- Matsuura M, Noah JW, Lambowitz AM. Mechanism of maturase-promoted group II intron splicing. *EMBO J.* 2001 Dec 17;20(24):7259-70.
- Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev.* 1997 Nov 1;11(21):2910-24.
- McKenna SA, Kim I, Puglisi EV, Lindhout DA, Aitken CE, Marshall RA, Puglisi JD. Purification and characterization of transcribed RNAs using gel filtration chromatography. *Nat Protoc.* 2007;2(12):3270-7.
- Michel F, Costa M, Westhof E. The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci.* 2009 Apr;34(4):189-99.
- Mitchell M, Gillis A, Futahashi M, Fujiwara H, Skordalakes E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol.* 2010 Apr;17(4):513-8.
- Mohr G, Ghanem E, Lambowitz AM. Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol.* 2010 Jun 8;8(6):e1000391.
- Mohr G, Perlman PS, Lambowitz AM. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res.* 1993 Nov 11;21(22):4991-7.
- Mohr G, Smith D, Belfort M, Lambowitz AM. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev.* 2000 Mar 1;14(5):559-73.
- Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, King O, Polioudakis D, Iyer VR, Hunicke-Smith S, Swamy S, Kuersten S, Lambowitz AM. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA.* 2013 Jul;19(7):958-70.
- Moon AF, Mueller GA, Zhong X, Pedersen LC. A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. *Protein Sci.* 2010 May;19(5):901-13.

- Moretz SE, Lampson BC. A group IIC-type intron interrupts the rRNA methylase gene of *Geobacillus stearothermophilus* strain 10. *J Bacteriol.* 2010 Oct;192(19):5245-8.
- Nallamsetty S, Waugh DS. Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expr Purif.* 2006 Jan;45(1):175-82.
- Perutka J, Wang W, Goerlitz D, Lambowitz AM. Use of computer-designed group II introns to disrupt *Escherichia coli* DExH/D-box protein and DNA helicase genes. *J Mol Biol.* 2004 Feb 13;336(2):421-39.
- Petoukhov MV, Konarev PV, Kikhney AG, Svergun DI. ATSAS 2.1 – towards automated web-based supported small-angle scattering data analysis. *J Appl Cryst.* 2007 40 s223-s228.
- Qin Y. Unpublished work.
- Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* 2010 Apr;5(4):725-38.
- Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry.* 1999 Jul 13;38(28):9069-83.
- San Filippo J, Lambowitz AM. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol.* 2002 Dec 13;324(5):933-51.
- Slabinski L, Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics.* 2007 Dec 15;23(24):3403-5.
- Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B. Crystal structures of fusion proteins with large-affinity tags. *Protein Sci.* 2003 Jul;12(7):1313-22.
- Svergun DI. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys J.* 1999 Jun;76(6):2879-86. Erratum in: *Biophys J.* 1999 Nov;77(5):2896.
- Svergun DI, Petoukhov MV, Koch MH. Determination of domain structure of proteins from X-ray solution scattering. *Biophys J.* 2001 Jun;80(6):2946-53.
- Vellore J, Moretz SE, Lampson BC. A group II intron-type open reading frame from the thermophile *Bacillus* (*Geobacillus*) *stearothermophilus* encodes a heat-stable reverse transcriptase. *Appl Environ Microbiol.* 2004 Dec;70(12):7140-7.
- Volkov VV, Svergun DI. Uniqueness of ab initio shape determination in small-angle scattering. *J Appl Cryst.* 2003 36 860-864.

- Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell*. 1999 Aug;4(2):239-50.
- Yao J, Truong DM, Lambowitz AM. Genetic and biochemical assays reveal a key role for replication restart proteins in group II intron retrohoming. *PLoS Genet*. 2013 Apr;9(4):e1003469.
- Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008 Jan 23;9:40.